基于网络用户评论的评分预测模型研究*

张红丽　刘济郢　杨斯楠　徐　健
(中山大学资讯管理学院　广州　510006)

摘要:【目的】通过网络用户评论, 为评论网站构建有效的评分预测机制。【方法】提出基于网络用户评论的评分预测模型, 该模型包括 4 个模块: 网络用户评论获取模块、预测变量获取模块、预测分析模块以及预测结果评价模块。抓取 30 部不同类型的电影评论数据, 27 部用于构建模型, 3 部用于检验模型。【结果】使用逐步回归方法筛选出变量: 参与评分人数、参与评论人数、想要观看人数和电影正向评论情感均值, 构建评分预测模型。使用 3 部电影验证, 预测评分与 IMDb 评分相差最大值为 0.0644, 最小值为 0.0227。【局限】在数据样本量、情感特征提取精度、模型普适性验证等方面有待进一步提升。【结论】该模型能够依据用户评论对评分进行有效预测, 在网络水军探测方面也能发挥一定的作用。

关键词: 评分预测　情感分析　回归分析　电影评分　网络水军探测
分类号: G350

chinaXiv:201712.01599v1

1　引　言

随着 Web2.0 的发展, 每一位网络用户都可以通过互联网发表个人对产品的观点并为产品打分, 专门的产品评分网站也应运而生。同时, 越来越多的消费者将评分网站上的用户评分作为消费决策的重要参考。但由于信息发布的门槛降低, 评分网站上的评分易受到非正常手段干扰, 面对评分网站上纷繁的产品宣传和产品评价, 如何从网络中识别真实的产品评价及评分成为网民们关注的问题。如今评分网站在引导消费上起到极其关键的作用, 但是其存在两个问题使得产品的真实性大打折扣: 一是消费者评论具有混杂性, 二是用户恶意刷分行为影响了产品的真实评分。普通用户只能通过网络评分辨别产品的优劣, 而一个不具有公信力的评分很大程度上会误导用户判断。另外, 网络评分在产品发布之后一段时间才趋于稳定, 存在滞后性的特点。

针对上述评分网站的问题, 本文通过选取网络用户评论的相关指标, 提出一种基于网络用户评论的评分预测模型。由于网络评论中包含用户对产品的意见和情感倾向, 因此, 基于用户的评论内容, 利用情感分析技术分析评论文本的情感倾向性, 将情感指标作为辅助预测指标, 以提高模型的预测效果。对于个人, 可以通过评分预测模型得到更客观公正的评分, 为消费决策提供建议; 对于商家, 可以收到最真实的使用反馈, 以改进产品质量; 对于网站管理方, 可以用来探测评分异常值存在, 及时发现"网络水军"[1], 维护网站正常运营。

2　相关研究

目前对网络用户评论的相关研究已经取得了较多的成果, 主要研究方向集中在评论的有用性、评论对产品销量的影响和评论文本挖掘三个方面。

(1) 评论的有用性是指用户产生的能够帮助潜在消费者购买决策的产品评价[2]。只有消费者认为有用的评价才具有实际价值, 研究者主要从评论内容和评论用户的角度对评论的有用性进行探索。Chen 等[3]抓取亚马逊网站用户评论数据, 提出网络用户评论的有

通讯作者: 徐健, ORCID: 0000-0003-4886-4708, E-mail: issxj@mail.sysu.edu.cn。
*本文系国家社会科学基金项目"用户评论情感分析及其在竞争情报服务中的应用研究"(项目编号: 11CTQ022)和广东省科技专项"基于内容的科技文献分析服务平台"(项目编号: 2016B030303003)的研究成果之一。

ChinaXiv合作期刊

chinaXiv:201712.01599v1

用性与评论用户、评论效用和评论获支持数存在较强的关联性。吴江等[4]从评论信息的相关性、及时性、客观性、真实性 4 个维度出发，构建评论有用性影响因素模型。Kuan 等[5]利用亚马逊评论数据探索出评论语句的长度、可读性程度、情感极性、评论用户的信誉对评论的有用性具有影响。

(2) 评论对产品销量的影响涉及的产品领域众多，主要包括电子产品类、音像图书类、旅游酒店类、电影类等。王文君等[6]通过对在线手机评论研究发现，评论长度、评论时效性、评论数量、负面评论和产品价格对在线手机销量有显著性影响。龚诗阳等[7]分析了当当网上的图书评论，研究显示评论数量对图书销量有正向影响。评论数量对销量的影响程度随着图书上线的时间变长而减弱。Torres 等[8]研究美国 178 家酒店在 TripAdvisor 上的评分排名与在线评论数量对酒店在线交易产生的影响，分析发现评论数量和评分排名对酒店在线预订交易具有积极影响。Chintagunta 等[9]测量了评论效用、评论数量对电影票房的影响。

(3) 评论文本挖掘主要包括产品特征挖掘和用户情感的判断。对评论中产品特征的挖掘是从产品自身的角度进行分析，Liu 等[10]首先提出应用关联规则分类方法提取英文评论中的产品特征。杜思奇等[11]引入汉语组块分析，结合支持向量机、Apriori 算法获取频繁项集、TF-IDF 停用词过滤实现评论文本中产品特征的提取。用户情感的判断主要通过挖掘用户网络评价的情感倾向分析用户对评价对象的褒贬态度。单晓红等[12]采用情感分析方法对苹果手机用户的网络评论进行分析，为用户购买决策提供依据。吴维芳等[13]利用 Word2Vec 对 TripAdvisor 酒店评论进行特征抽取和降维，结合情感分析技术，构建计量经济模型分析酒店特征评价与用户满意度的关系。

另外，在评分预测方面，马春平等[14]提出一种基于词向量的方法挖掘用户评论信息，并结合协同过滤的方法设计新的推荐算法，该算法有效地提高了推荐系统的评分预测性能。Kamath 等[15]利用 MG- LDA[16]算法对评论进行主题分析生成主题词表，利用主题词表将用户评论表示成特征向量，利用机器学习算法建模进行评分预测。马松岳等[17]对豆瓣电影的用户评价进行情感分析得到综合情绪值，发现评论评价的综合情绪值与打分评价相关性较高，根据评论评价构建预测打分模型。但该模型变量只涉及综合情绪值和评论总数，没有考虑评论的其他因素。

综上所述，目前虽然有很多关于网络用户评论的研究，但研究主要集中于评论效用和挖掘技术方面。在评分预测方面，结合情感分析，并用于评论分数预测方面的相关研究较少。本文在网络用户评论相关变量基础上，引入情感特征因素作为辅助预测变量，提出基于网络用户评论的评分预测模型，旨在利用情感分析和回归分析手段实现对产品评分网站客观评分的有效预测。

3 基于网络用户评论的评分预测模型设计

本文提出一种基于网络用户评论的评分预测模型，预测评分网站中产品的客观评分。借助情感分析的手段，提取用户语料中的情感特征，使之成为辅助预测指标，并寻找行业内最客观公正的评分作为预测对比变量。同时结合相关联的预测指标以及情感分析指标作为自变量，通过回归分析构建评分预测模型。该模型主要由 4 个部分构成：网络用户评论获取模块、预测变量获取模块、预测分析模块以及预测结果评价模块，如图 1 所示。

(1) 网络用户评论获取模块主要包括网络评论来源的筛选以及网络评论数据的获取。质量高的数据源有助于模型的有效建立，选定具有代表性的网站作为网络评论数据源[18]；选取行业客观评分数据来源；采集所需数据并存储在数据库中。

(2) 预测变量获取模块主要包括网络用户评论相关预测指标和情感特征指标。获取网络用户评论相关预测指标，对数量级大的变量进行对数缩放操作，防止数据的量级差距导致模型失真。情感特征指标提取包括数据清洗、中文分词、去停用词以及情感量化[19]。对网络用语化且非结构化的网络用户评论进行数据清洗，剔除评论中的网络链接、表情等非规范信息，只保留文本内容；进行文本分词和去停用词处理，减少情感量化的计算量；通过情感值计算的方式对语料数据进行量化。

(3) 预测分析模块主要针对预测变量，采用多元线性回归分析方法构建预测模型[20]，并对模型进行结果分析。若 P 值不显著，则采用不同的回归分析方法筛选变量，重构模型，观察各个变量 P 值是否显著(小

于 0.05)，倘若不显著说明模型建立失败。若 P 值显著，再对 R 方(R-square)和调整 R 方(Adjusted R-square)进行比较，选取值较高的回归模型，该数值越大，预测值与实际值越接近。

(4) 预测结果评价模块主要包括对回归模型的预测结果进行可视化解析，通过拟合预测分数与客观评分，观察预测效果。倘若预测中出现异常值和不显著的变量，分析其差异性的缘由，进行剔除后，重新构建回归方程，并采用预测数据检验模型的实际预测效果，以证明预测模型的有效性。

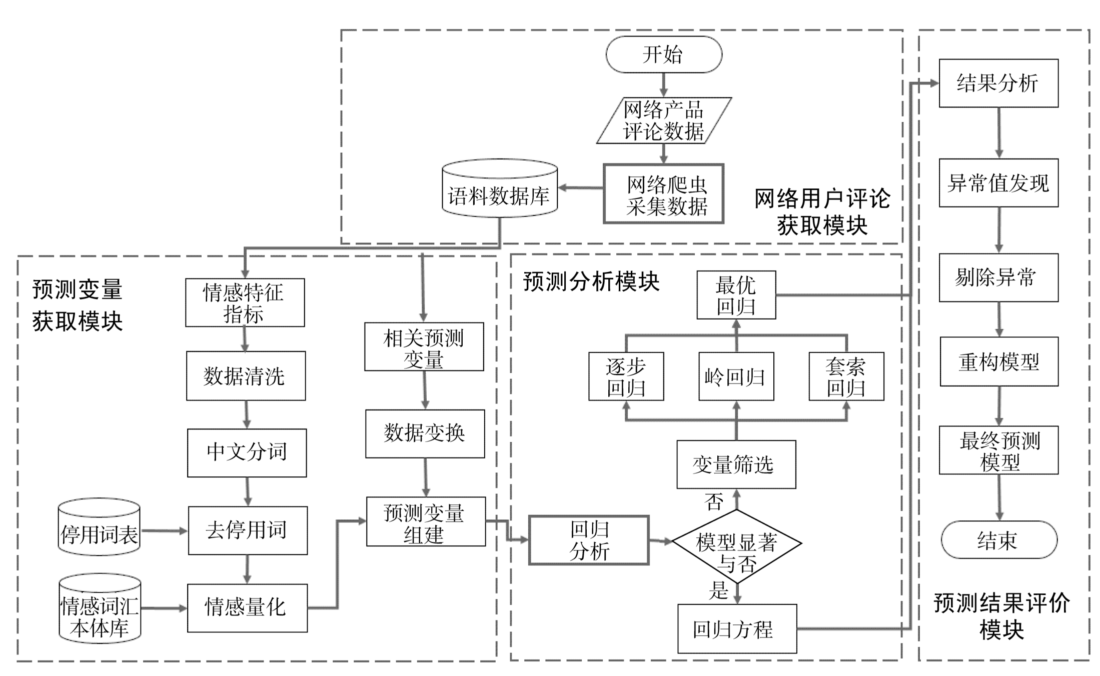


图 1　评分预测模型流程

4　模型验证与评估

为验证评分预测模型的有效性，以电影评分网站为例，通过网络用户评论预测模型来预测电影评分。

4.1　实验数据来源选取与采集

(1) 豆瓣电影影评数据源

豆瓣电影是国内热门的电影评分网站，收录了十分齐全的国内外电影数据，用户数量及电影评论数据量巨大，是一个理想的网络评论源。豆瓣的影评主要以两种形式存在：短评和长评。短评字数限制在 140 字以内，主要是豆瓣用户对于电影较为宏观或者某个方面的评价。长评多为篇幅型影评内容，内容繁杂，很多电影之外的内容，例如有些会介绍拍摄过程、拍摄手法或者演职人员等。因此，本文选取豆瓣电影评分网站的短评作为网络用户评论语料。

以近年来的电影为样本，为保障数据的多样化，选取时尽量兼顾电影上映月份和不同类型的电影题材，如动作类、喜剧类、科幻类等。共计选择 30 部电影，部分电影如表 1 所示。

表 1　电影样本(部分)

编号	电影名称	国内上映日期	类型	制作地区
1	小时代 4	2015/7/9	爱情、剧情、青春	中国内地、中国台湾
2	小时代 2	2013/8/8	青春、剧情、爱情	中国内地、中国台湾
3	恶棍天使	2015/12/24	喜剧、荒诞、爱情	中国
4	万物生长	2015/4/17	爱情、剧情、校园	中国
5	捉妖记	2015/7/16	剧情、喜剧、奇幻	中国
6	湄公河行动	2016/9/30	动作、警匪	中国
7	驴得水	2016/10/28	喜剧、剧情	中国
8	功夫熊猫 3	2016/1/29	动画，喜剧、动作	美国、中国
9	百鸟朝凤	2016/5/6	剧情、文化	中国
10	七月与安生	2016/9/14	剧情、爱情、青春	中国

(2) 客观评分数据源

互联网电影数据库(IMDb)是目前信息量较大、使用人数较多、影响范围较广、影响力较大的电影网站

chinaXiv:201712.01599v1

ChinaXiv合作期刊

chinaXiv:201712.01599v1

之一[21]。IMDb 的影片得分采取统计学的计算方法，并结合部分专家的评分意见，保障电影的评分不受极端行为的影响。为保障电影评分的客观性，本文选取 IMDb 的评分系统作为评分预测模型的客观评分来源。

(3) 电影影评时间区间选取

由于电影的影评数据时间轴较长，通过观察电影影评趋于稳定状态的时长，确定选取数据的时间区间。一般来说，多数电影的上映期限为一个月。选取不同类型的电影《百鸟朝凤》、《七月与安生》、《功夫熊猫 3》，对其上映后获取的数据量进行分析，如图 2 所示。

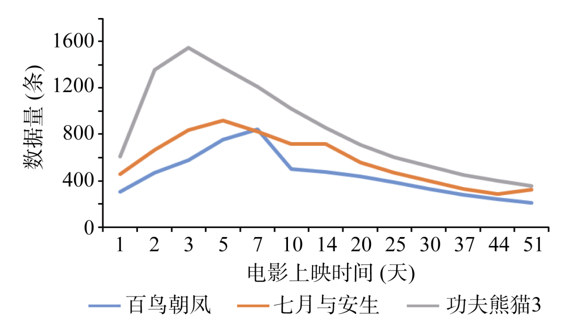


图 2　《百鸟朝凤》、《七月与安生》、《功夫熊猫 3》豆瓣影评增长趋势

从图 2 可知，三部电影的评论数据在上映后一周达到顶峰，在 30 天后评论数据波动不再明显，并趋于稳定。此外，在分析三部电影的豆瓣电影短评情感倾向性方面出现类似现象，如《七月与安生》，正向评论情感值和负向评论情感值在第一周内波动较为明显，随着上映时间的推移，情感值均在 30 天左右逐渐趋于稳定。电影上映第 30 天，情感值均值稳定在 1.7 左右，浮动很小，如图 3 所示。

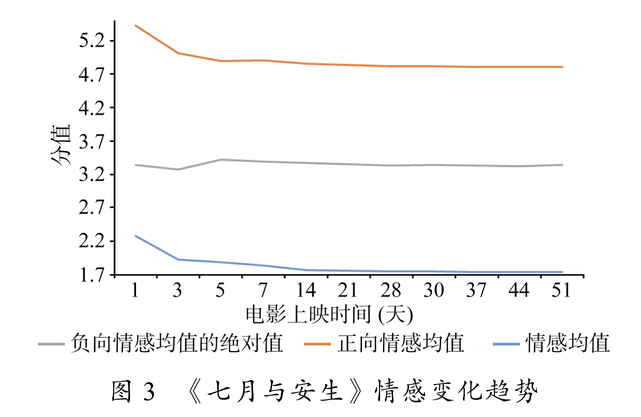


图 3　《七月与安生》情感变化趋势

综上所述，若评论数据的波动性太大，会导致情感量化结果出现偏差，实际预测结果失真。因此，在模型构建时，要选取能够反映稳定情感的数据源。本实验中选取电影上映之后 30 天内的豆瓣电影评论数据作为语料数据来源。

使用爬虫软件“集搜客”[22]抓取豆瓣电影影评(包括短评用户名、短评内容、评论时间、获得支持数及评分数)作为实验数据集，选用 IMDb 为客观评分来源。共抓取 30 部电影 1 469 660 条电影短评，数据去重后选取电影上映后 30 天内的短评数据，共计 513 788 条。

4.2　预测变量获取

(1) 网络用户评论相关预测指标

网络评论预测变量通过豆瓣电影页面相关数据选取：评分人数(*criticNum*)指参与该电影评分的用户数；参与评论人数(*commentNum*)指参与该电影的文字评论的用户数；标记看过人数(*watchedNum*)指已经看过该部电影的用户数；想看的人数(*desireNum*)指在豆瓣上标记了对这部电影感兴趣或者想要观看的用户数。其中，开始选择想看的用户，看过电影后改为看过，将不再在想看那组，即两组互斥。根据所获得数据延展出两个变量：参与电影评论的比例(*comment Ratio*)和想看人数比例(*desireRatio*)，计算方法如公式(1)和公式(2)所示。

$$commentRatio = \frac{commentNum}{watchedNum} \tag{1}$$

$$desireRatio = \frac{desireNum}{(desireNum + watchedNum)} \tag{2}$$

commentRatio 是评论人数在看过人数中的占比，表示想表达对电影观点的影迷占比情况。很多影迷在未观看电影前先对电影进行标记，表明对电影有极大的兴趣，*desireRatio* 表示想看人数占想看人数和已看过人数之和的比例，可反映对电影的喜爱程度。由于获取的数据量级比较大，为避免模型失真，本文采用底数为 10 的对数缩放方法对数据进行变换，例如 *criticNum* 变换后的变量名为 *LcriticNum*。

(2) 情感特征指标

本文情感量化采用基于情感词典的方式，使用大连理工大学的情感词汇本体库[23]。本体库中词汇的情感强度 1、3、5、7、9 级别分别对应 1、2、3、4、5 分，正向情感为正数，负向情感为负数，中性词为零。例如，“阻力”在本体库中被标注为负向情感词并且情感强度为 3，其对应的情感分数为–2 分。*sentimentScore*

代表某条评论的情感分数，i 代表评论中正向词的序列数，P_i 代表该词对应的正向情感分数。j 代表评论中负向词的序列数，N_j 代表该词对应的负向情感分数，假设评论中共有 n 个正向情感词，m 个负向情感词，情感分数计算如公式(3)所示。

$$sentimentScore = \sum_{i=1}^{n} P_i + \sum_{i=1}^{m} N_j \quad (3)$$

对30天的电影评论数据的情感进行量化，并求出情感均值(*sentimentmeanScore*)。为更好地表达电影的情感倾向，在情感均值的基础上，计算正向情感均值(*posmeanScore*)和负向情感均值(*negmeanScore*)。正向情感均值为30天电影评分数据正向评价的算术平均值，负向情感均值为30天电影评分数据负向评价的算术平均值。i、j、k 分别指代某条评论数据; a 表示正向评论数量; b 表示负向评论数量; n 指总数量，即 $n=a+b$; $pos(i)$指第 i 条评论的正向情感值; $neg(j)$指第 j 条评论的负向情感值; $sentimentScore(k)$指第 k 条评论的情感值。计算如公式(4)–公式(6)所示。

$$posmeanScore = \frac{\sum_{i=1}^{a} pos(i)}{a} \quad (4)$$

$$negmeanScore = \frac{\sum_{j=1}^{b} neg(j)}{b} \quad (5)$$

$$sentimentmeanScore = \frac{\sum_{k=1}^{n} sentimentScore(k)}{n} \quad (6)$$

提取完所有电影的情感特征后，组建出所有的预测变量及含义(见表 2)，并归纳整理变量数据(部分数据见表 3)。

表 2　预测变量及含义

预测变量名称	实际含义
LcriticNum	参与评分的人数以 10 为底对数值
LcommentNum	参与评论的人数以 10 为底对数值
LwatchedNum	已经看过的人数以 10 为底对数值
LdesireNum	想要观看的人数以 10 为底对数值
commentRatio	评论人数占评分人数的比例
desireRatio	想要观看人次占看过和想看人次的比例
sentimentmeanScore	电影评论情感均值
posmeanScore	电影正向评论情感均值
negmeanScore	电影负向评论情感均值
doubanScore	豆瓣电影评分

表 3　预测变量值表(部分)

编号	电影名称	Lcritic Num	Lcomment Num	Lwatched Num	Ldesire Num	comment Ratio	desireRatio	sentiment meanScore	posmean Score	negmean Score	douban Score
1	小时代 4	4.9019	4.5759	4.9563	3.9654	0.4720	0.0927	0.6022	4.3345	–3.7442	4.6
2	小时代 2	5.1045	4.7196	5.1774	3.8624	0.4121	0.0462	0.6174	4.2995	–3.7318	5
3	恶棍天使	4.8992	4.6329	4.9357	3.8567	0.5416	0.0769	0.3044	4.1802	–3.6735	4
4	万物生长	4.9530	4.5765	5.0190	3.9803	0.4202	0.0838	0.5267	4.1363	–3.8332	5.9
5	捉妖记	5.3677	4.9937	5.4185	4.2924	0.4226	0.0696	1.2405	4.3430	–3.4054	6.8
6	湄公河行动	5.3412	5.0007	5.3659	4.5103	0.4565	0.1224	1.4745	4.6532	–3.5063	8.1
7	驴得水	5.1235	4.7927	5.1492	4.4252	0.4668	0.1588	0.4241	4.3093	–4.1345	8.3
8	功夫熊猫 3	5.1937	4.7917	5.2385	4.0827	0.3962	0.0653	1.7018	4.6260	–3.0234	7.7
9	百鸟朝凤	4.9233	4.5974	4.9611	4.3204	0.4722	0.1861	2.1067	5.5629	–3.3765	8
10	七月与安生	5.2082	4.8858	5.2441	4.2882	0.4760	0.0997	1.7458	4.8169	–3.3355	7.6

4.3　预测分析

回归分析方法可以用来判别客观事物数量的依存关系，可以用来处理多个变量之间相互关系。回归分析是研究相关关系的一种数学方法，是寻找不完全确定的变量间的数学关系式并进行统计推断的一种方法[24]。常见的回归预测有多元线性回归(Multiple Regression)[25]、逐步回归(Stepwise Regression)[26]、岭回归(Ridge Regression)[27]、套索回归(Lasso Regression)[28]等方法。

针对上述的数据变量，分别使用多元线性回归、

chinaXiv:201712.01599v1

ChinaXiv合作期刊

chinaXiv:201712.01599v1

逐步回归、岭回归以及套索回归方法对模型进行变量选择，构建预测模型，确定最优回归方程。采用 30 部电影中 27 部电影数据作为模型构建数据，3 部电影作为检验数据。

由于数据涉及到多个变量，但无法判断各变量在模型中关联程度的大小，因此使用多元线性回归，观察各变量 P 值的大小，结果如表 4 所示。

表 4　多元线性回归各变量 P 值

变量名	P 值
LcriticNum	0.142
LcommentNum	0.217
LwatchedNum	0.304
LdesireNum	0.151
commentRatio	0.359
desireRatio	0.308
sentimentmeanScore	0.824
posmeanScore	0.427
negmeanScore	0.820

当所有变量加入到多元线性回归时，最大值 *wacthedNum* 为 0.75，远大于 0.05；最小值 *LcriticNum* 也达到 0.142，所有变量的 P 值均大于 0.05。构建多元线性回归模型失败，需要对变量进行筛选。

使用逐步回归、岭回归以及套索回归分别对模型进行变量选取，并观察各个变量的 P 值，如表 5 所示。

表 5　三种回归方法各变量 P 值

回归方法	变量名	P 值
逐步回归	*LcriticNum*	0.0320
	LcommentNum	0.0046
	LwacthedNum	0.0728
	LdesireNum	0.0027
	posmeanScore	0.0020
岭回归	*LdesireNum*	0.0001
	commentRatio	0.0336
	posmeanScore	0.0020
套索回归	*LdesireNum*	0.0001
	sentimentmeanScore	0.0003

通过对比逐步回归、岭回归、套索回归三种回归分析的统计量来分析上述三种模型的实际预测效果，各 P 值均表示模型显著，进一步探索三种模型 R 方和调整 R 方，如图 4 所示。

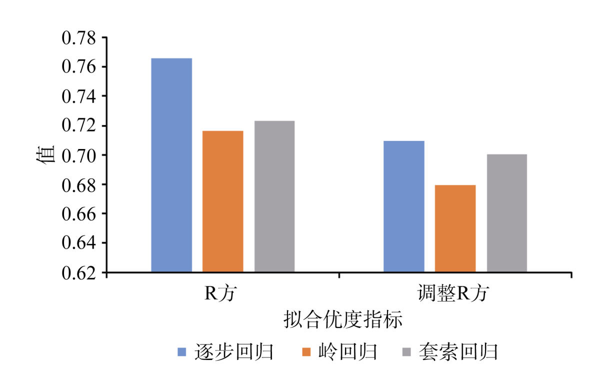


图 4　逐步回归、岭回归、套索回归模型统计量对比

岭回归在两个指标上都是最弱的，且调整 R 方的值与逐步回归、套索回归的差距非常大。对于调整 R 方，逐步回归的值和套索回归的值相对较高，但是逐步回归的 R 方值最高，达到 0.7656，拟合效果较佳。因此，最优选择为逐步回归方法构建的回归方程，如公式(7)所示。

$$Y=-12.9328+35.7904\times LcriticNum-11.5032\times LcommentNum-24.6262\times LwacthedNum+2.9563\times LdesireNum+1.2417\times posmeanScore \quad (7)$$

4.4　预测结果评价

预测分析后，还需对得到的预测模型进行评价。若出现异常值，需分析原因，剔除异常值后重构模型，并用检验数据对模型进行检验。

(1) 预测结果分析

使用最优回归方程公式(7)对各电影评分进行预测，结果如图 5 所示。

通过拟合 IMDb 分数与评分预测值，可以发现大部分电影之间的差距很小，误差值在很小的范围内，说明预测模型整体上是有效的。其中有几部电影差距较为明显，例如《小时代 2》和《小时代 4》预测分数明显大于其 IMDb 分数。

(2) 异常值发现

从模型的预测结果来看，正常电影评分预测值和 IMDb 值之间差距往往不超过 1 分，本文定义预测值与 IMDb 值差距超过 1 分的为异常值，如图 6 所示。

从图 6 可知，拟合正常情况下的电影如《明日边缘》、《火星救援》，预测值与 IMDb 分数的差距很小。而《小时代 2》、《小时代 4》 预测值与 IMDb 值差距超过 1 分，甚至 2 分。可以判断这两部电影的评论数

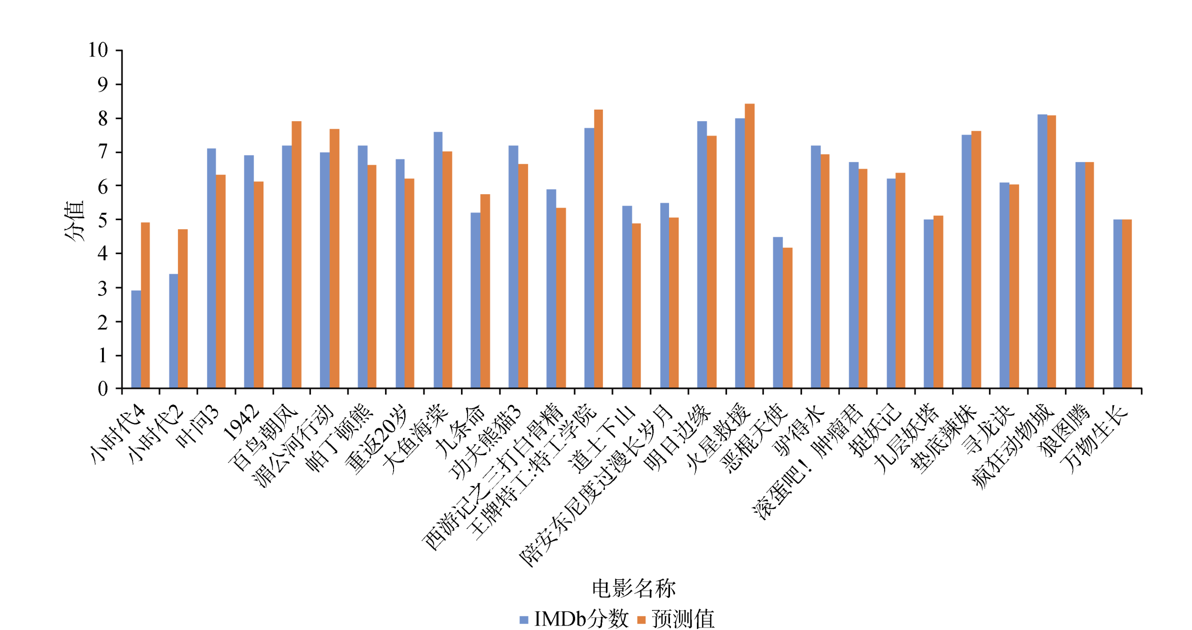


图 5　逐步回归构建模型预测评分与实际评分的直方图

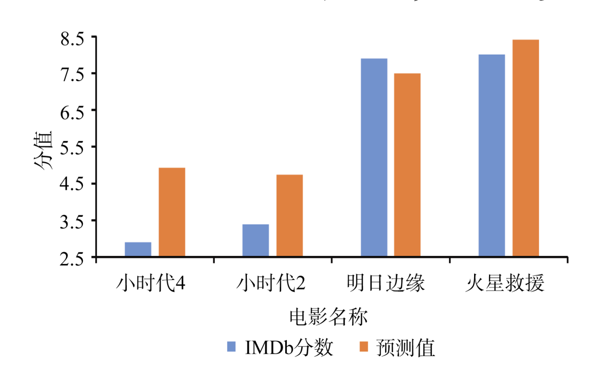


图 6　异常值和正常值拟合效果对比

据情感倾向具有非真实性。通过查阅新闻和文献证实两部电影确实存在刷分行为，说明本模型不仅具有评分预测的作用，在“网络水军”探测方面也发挥一定的作用。

(3) 剔除异常值并重构模型

为避免异常值对模型的干扰，剔除《小时代 2》和《小时代 4》的数据，利用逐步回归的方法重新构建预测方程。此外，新的回归模型剔除了 P 值略高的 *LwatchedNum*，仅使用 *LcriticNum*、*LcommentNum*、*LdesireNum* 以及 *posmeanScore*，这些变量的 P 值都具有极高的显著性，如表 6 所示，构建回归方程如公式(8)所示。

$$Y = -11.1349 + 7.4531 \times LcriticNum - 7.4636 \times LcommentNum + 2.3371 \times LdesireNum + 1.1499 \times posmeanScore \quad (8)$$

表 6　剔除异常值后逐步回归变量 P 值

变量名	P 值
LcriticNum	0.0003
LcommentNum	0.0004
LdesireNum	0.0002
posmeanScore	0.0001

新的回归分析结果的统计量如图 7 所示，剔除异常值后的 R 方和调整 R 方明显提升，R 方的值达到 0.8572，调整 R 方的值达到 0.8287，模型的预测效果较好。

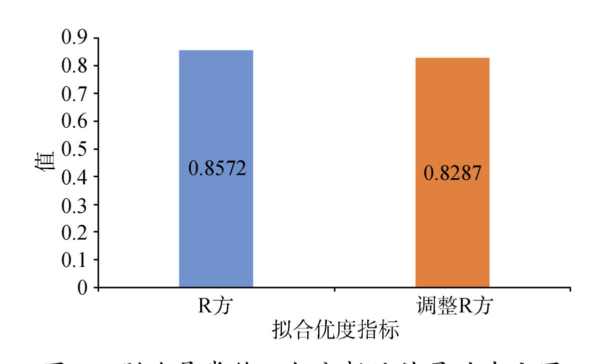


图 7　剔除异常值回归分析统计量的直方图

对比新模型拟合的预测值与 IMDb 分数如图 8 所示，可以明显看出，各个电影的预测值和 IMDb 值之间差距较小，最大差距的为《叶问 3》，差值为 0.7 分；最小差距的为《垫底辣妹》，差值仅为 0.05 分。因此，公式(8)具有较好的预测效果，根据方程中的变量要

chinaXiv:201712.01599v1

ChinaXiv合作期刊

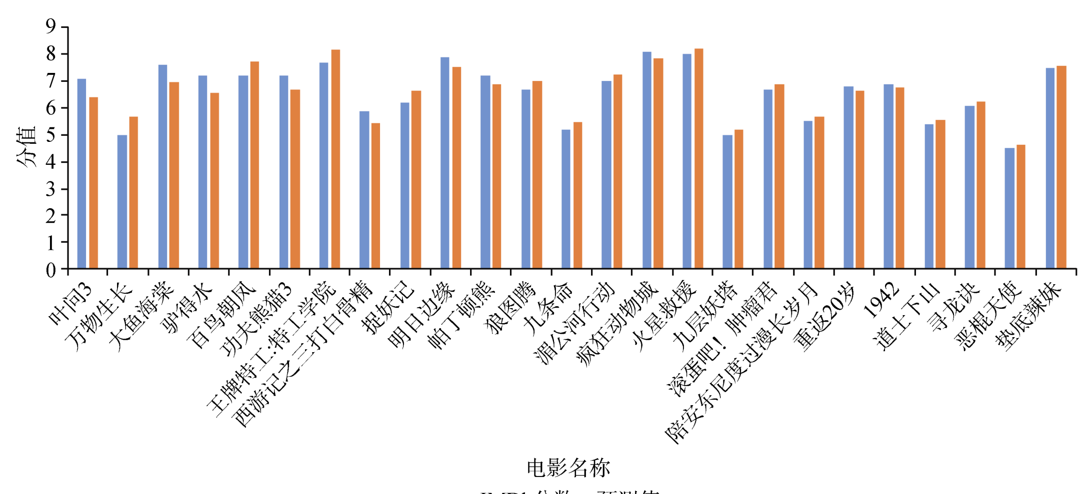


图 8　剔除异常值后回归构建模型预测评分与 IMDb 分数的直方图

求，仅需要其电影的 *LcriticNum*、*LcommentNum*、*LdesireNum* 和 *posmeanScore* 就可以对电影的客观评分进行预测。

(4) 模型检验

为了检验模型实际效果，使用预留的三部电影数据进行评分预测，分别为《心迷宫》、《七月与安生》以及《我的少女时代》，相关变量如表 7 所示。

表 7　评分预测模型检验数据

电影名称	*LcriticNum*	*LcommentNum*	*LdesireNum*	*posmeanNum*
心迷宫	5.1247	4.7244	4.6835	4.9646
七月与安生	5.2082	4.8858	4.2882	4.8169
我的少女时代	5.3919	5.0585	4.4110	4.8415

利用公式(8)对三部电影的评分进行预测，结果如图 9 所示。

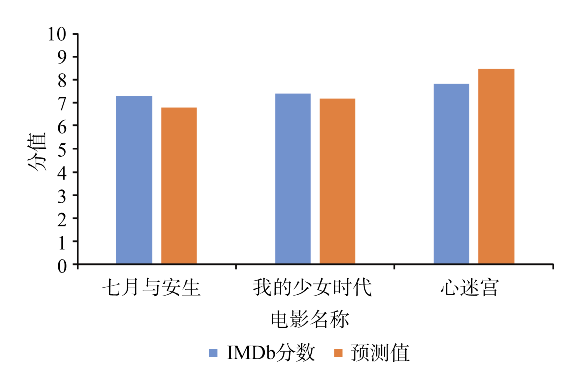


图 9　模型实际预测效果

可以看出三部电影的评分预测值与 IMDb 实际值都很接近且误差很小，《七月与安生》的误差为 0.0522，《我的少女时代》的误差为 0.0227，《心迷宫》的误差为 0.0644，因此，模型的实际预测效果较理想。

5　结　语

互联网环境下，评分网站不容忽视，一方面为潜在消费者选购商品提供决策参考，另一方面为商家提供商机。评分网站由于开放性导致产品评分失真，客观的评分网站需求愈发迫切。本文提出基于网络用户评论的评分预测模型来预测客观评分，该模型主要包括网络用户评论获取、预测变量获取、预测分析以及预测结果评价 4 个模块。为验证评分预测模型的有效性，以"豆瓣电影"的评论内容作为语料来源，以 IMDb 作为客观评分来源。对近年来 30 部不同类型的电影影评进行实证研究，结果显示，在评分预测模型中，电影上映 30 天时的评论数据稳定性最高，最适合用作预测数据源。在回归分析中，逐步回归方式筛选出变量构建的回归方程预测效果最优。在预测分数和 IMDb 分数拟合阶段，发现异常值，说明本模型不仅具有评分预测的作用，在"网络水军"探测方面也有一定的作用。剔除异常值后，仅需要其电影的 *LcriticNum*、*LcommentNum*、*LdesireNum* 和 *posmeanScore* 变量就可以对电影的客观评分进行预测，重构模型之后利用三部电影对模型评分预测效果进行检验，预测评分效果较佳。

本文存在以下不足之处：数据样本量较少，可考虑通过增加数据量优化模型预测效果；此外，在情感

chinaXiv:201712.01599v1

ChinaXiv合作期刊

分析技术方面，主要是基于词典技术进行情感特征提取，未来可尝试结合机器学习方法或者其他前沿的情感分析技术进一步精确提取情感特征；除了以电影评分网站作为实例外，可选取其他类型评分网站的数据进行实证研究，以验证模型的普适性。

参考文献：

[1] 楼旭东，刘萍. "网络水军"的传播学分析[J]. 当代传播，2011(4): 76-77. (Lou Xudong, Liu Ping. A Communicational Analysis of the "Water-forces in the Network" [J]. Contemporary Communication, 2011(4): 76-77.)

[2] Mudambi S M, Schuff D. What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com[J]. MIS Quarterly, 2010, 34(1): 185-200.

[3] Chen Y, Chai Y, Liu Y, et al. Analysis of Review Helpfulness Based on Consumer Perspective [J]. Tsinghua Science & Technology, 2015, 20(3): 293-305.

[4] 吴江，刘弯弯. 基于信息采纳理论的在线商品评论有用性影响因素研究[J]. 信息资源管理学报，2017, 7(1): 47-55. (Wu Jiang, Liu Wanwan. A Research of Factors Affecting the Perceived Helpfulness of Online Product Based on the Information Adoption Theory [J]. Journal of Information Resources Management, 2017, 7(1): 47-55.)

[5] Kuan K K, Hui K, Prasarnphanich P, et al. What Makes a Review Voted? An Empirical Investigation of Review Voting in Online Review Systems[J]. Journal of the Association for Information Systems, 2015, 16(1): 48-71.

[6] 王文君，张静中. 电子商务网站在线评论对手机销量影响的实证研究[J]. 河北工业科技，2016, 33(3): 188-193. (Wang Wenjun, Zhang Jingzhong. An Empirical Study of the Impact of Online Reviews on Mobile Phone Sales in E-commerce[J]. Hebei Journal of Industrial Science and Technology, 2016, 33(3): 188-193.)

[7] 龚诗阳，刘霞，赵平. 线上消费者评论如何影响产品销量?——基于在线图书评论的实证研究[J]. 中国软科学，2013(6): 171-183. (Gong Shiyang, Liu Xia, Zhao Ping. How do Online Consumer Reviews Influence Product Sales? —An Empirical Study Based on Online Book Reviews.[J] China Soft Science, 2013(6): 171-183.)

[8] Torres E N, Singh D, Robertson-Ring A. Consumer Reviews and the Creation of Booking Transaction Value: Lessons from the Hotel Industry [J]. International Journal of Hospitality Management, 2015, 50: 77-83.

[9] Chintagunta P K, Gopinath S, Venkataraman S, et al. The Effects of Online User Reviews on Movie Box Office Performance: Accounting for Sequential Rollout and Aggregation Across Local Markets[J]. Marketing Science, 2010, 29(5): 944-957.

[10] Liu B, Hu M, Cheng J. Opinion Observer: Analyzing and Comparing Opinions on the Web[C]//Proceedings of the 14th International Conference on World Wide Web, Chiba, Japan. New York, USA: ACM, 2005: 342-351.

[11] 杜思奇，李红莲，吕学强. 汉语组块分析在产品特征提取中的应用研究[J]. 现代图书情报技术，2015(9): 26-30. (Du Siqi, Li Honglian, Lv Xueqiang. Research of Chinese Chunk Parsing in Application of the Product Feature Extraction[J]. New Technology of Library and Information Service, 2015(9): 26-30.)

[12] 单晓红，杨柳. 网络产品评论挖掘研究[J]. 计算机系统应用，2014, 23(2): 1-6. (Shan Xiaohong, Yang Liu. Research on Online Product Review Mining[J]. Computer Systems & Applications, 2014, 23(2): 1-6.)

[13] 吴维芳，高宝俊，杨海霞，等. 评论文本对酒店满意度的影响：基于情感分析的方法[J]. 数据分析与知识发现，2017, 1(3): 62-71. (Wu Weifang, Gao Baojun, Yang Haixia, et al. The Impacts of Reviews on Hotel Satisfaction: A Sentiment Analysis Method[J]. Data Analysis and Knowledge Discovery, 2017, 1(3): 62-71.)

[14] 马春平，陈文亮. 基于评论主题分析的评分预测方法研究[J]. 中文信息学报，2017, 31(2): 204-211. (Ma Chunping, Chen Wenliang. A Review Topic Analysis Method for Rating Prediction[J]. Journal of Chinese Information Processing, 2017, 31(2): 204-211.)

[15] Kamath R, Ochi M, Matsuo Y. Understanding Rating Behaviour and Predicting Ratings by Identifying Representative Users[OL]. arXiv PrePrint, arXiv: 1604.05468v1.

[16] Titov I, McDonald R. Modeling Online Reviews with Multi-grain Topic Models[C]// Proceedings of the 17th International Conference on World Wide Web. ACM, 2008: 111-120.

[17] 马松岳，许鑫. 基于评论情感分析的用户在线评价研究——以豆瓣网电影为例[J]. 图书情报工作，2016, 60(10): 95-102. (Ma Songyue, Xu Xin. Study on User Online Evaluation Based on Sentiment Analysis of Comments: Taking Douban.com Movie as an Example[J]. Library and Information Service, 2016, 60(10): 95-102.)

[18] 程翠琼，徐健. 面向网络游记时间特征的情感分析模型[J]. 数据分析与知识发现，2017, 1(2): 87-95. (Cheng Cuiqiong, Xu Jian. A Sentiment Analysis Model Based on Temporal

chinaXiv:201712.01599v1

ChinaXiv合作期刊

chinaXiv:201712.01599v1

Characteristics of Travel Blogs[J]. Data Analysis and Knowledge Discovery, 2017, 1(2): 87-95.)

[19] 吴应良，黄媛，王选飞. 在线中文用户评论研究综述：基于情感计算的视角[J]. 情报科学，2017，35(6): 159-163. (Wu Yingliang, Huang Yuan, Wang Xuanfei. Research on Online Users' Reviews in Chinese: Basing on the Perspective of Affective Computing[J]. Information Science, 2017, 35(6): 159-163.)

[20] 冷建飞，高旭，朱嘉平. 多元线性回归统计预测模型的应用[J]. 统计与决策，2016(7): 82-85. (Leng Jianfei, Gao Xu, Zhu Jiaping. Application of Multivariate Linear Regression Statistical Prediction Model [J]. Statistics and Decision, 2016(7): 82-85.)

[21] 王伟. 美国电影网站 IMDb 的榜单文化研究[D]. 长春：东北师范大学，2016. (Wang Wei. An Empirical Analysis of Factors Influencing the Helpfulness of Online Consumer Reviews[D]. Changchun: Northeast Normal University, 2016.)

[22] GooSeeker 集搜客网络爬虫，简单高效的网页采集器[EB/OL]. [2017-03-20]. http://www.gooseeker.com/. (GooSeeker Web Crawler, Simple and Efficient Web Collector[EB/OL]. [2017-03-20]. http: //www.gooseeker.com/.)

[23] 徐琳宏，林鸿飞，潘宇，等. 情感词汇本体的构造[J]. 情报学报，2008, 27(2): 180-185. (Xu Linhong, Lin Hongfei, Pan Yu, et al. Constructing the Affective Lexicon Ontology[J]. Journal of the China Society for Scientific and Technical Information, 2008, 27(2): 180-185.)

[24] Ray S. 7 Types of Regression Techniques You Should Know! [EB/OL]. [2017-03-20]. https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/.

[25] Abyaneh H Z. Evaluation of Multivariate Linear Regression and Artificial Neural Networks in Prediction of Water Quality Parameters[J/OL]. Iranian Journal of Environmental Health Science & Engineering, 2014. DOI: 10.1186/2052-336x-12-40.

[26] Yu T, Yu G, Li P Y, et al. Citation Impact Prediction for Scientific Papers Using Stepwise Regression Analysis[J]. Scientometrics, 2014, 101(2): 1233-1252.

[27] Wan S, Mak M, Kung S, et al. R3P-Loc: A Compact Multi-label Predictor Using Ridge Regression and Random Projection for Protein Subcellular Localization[J]. Journal of Theoretical Biology, 2014, 360: 34-45.

[28] Buccheri S, Capodanno D, Barbanti M, et al. A Risk Model for Prediction of 1-Year Mortality in Patients Undergoing MitraClip Implantation[J]. American Journal of Cardiology, 2017, 119(9): 1443-1449.

作者贡献声明:

徐健：提出研究思路，设计研究方案；
刘济郢：采集、清洗和分析数据，进行实验；
张红丽：论文起草；
张红丽，杨斯楠，徐健：论文最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据[1-3]见期刊网络版 http://www.infotech.ac.cn；支撑数据[4-7]由作者自存储, E-mail: issxj@mail.sysu.edu.cn。

[1] 张红丽，刘济郢，杨斯楠，徐健. movies.docx. 30 部电影选取完整表.

[2] 张红丽，刘济郢，杨斯楠，徐健. variable list.docx. 各电影相关数据一览表.

[3] 张红丽，刘济郢，杨斯楠，徐健. 停用词表.docx. 去除停用词用的停用词表.

[4] 张红丽，刘济郢，杨斯楠，徐健. 爬虫代码.docx. 集搜客豆瓣影评爬虫规则.

[5] 张红丽，刘济郢，杨斯楠，徐健. 分词代码.docx. Python 下结巴中文词源代码.

[6] 张红丽，刘济郢，杨斯楠，徐健. 情感值代码.docx. Python 下情感量化源代码.

[7] 张红丽，刘济郢，杨斯楠，徐健. 回归分析代码.docx. R 语言回归分析源代码.

收稿日期: 2017-05-31
收修改稿日期: 2017-07-19

ChinaXiv合作期刊

Predicting Online Users' Ratings with Comments

Zhang Hongli Liu Jiying Yang Sinan Xu Jian
(School of Information Management, Sun Yat-Sen University, Guangzhou 510006, China)

Abstract: [**Objective**] This study aims to build an effective prediction mechanism for online ratings, with the help of Web surfers' comments. [**Methods**] We proposed a model with the following modules: Web users'comment acquisition, predictive variable acquisition, prediction analysis and the prediction results evaluation. We retrieved 30 movies of different types and user's comments from the Web. 27 movies were used to build the model, which were then examined with the remaining movies. [**Results**] We employed the stepwise regression to select variables, which included the number of raters, the number of participants posting comments, the number of people who wanted to watch the moive and the sentiment value of the positive comments. The prediction results were quite close to the IMDb scores, and the maximum and the minimum differences were 0.0644 and 0.0227. [**Limitations**] The sample size, the accuracy of sentiment features, and compatibility of the model could be improved. [**Conclusions**] The proposed model effectively predicts movie scores and detects the "water army" online.

Keywords: Rating Prediction Sentiment Analysis Regression Analysis Movie Rating "Water Army" Detection

chinaXiv:201712.01599v1

人工智能有助于早期皮肤癌检测

滑铁卢大学和 Sunnybrook 研究所的研究人员开发了一项新技术，使用人工智能(AI)来辅助早期的黑色素瘤皮肤癌检测。该技术采用机器学习软件分析皮肤损伤的图像，并为医生提供黑色素瘤的生物标志物指示的客观数据。

该人工智能系统使用成千上万的皮肤图像及其相应的黑色素和血红蛋白水平进行训练，可以减少不必要的活检，大大节省了医疗成本。它能在医生采取更多的侵入性治疗行动之前，为医生提供病变特征的客观信息，以帮助他们排除黑色素瘤。

该技术最早将在 2018 年提供给医生使用。目前，皮肤病学家主要依靠皮肤病变的主观视觉检查来确定患者是否应该进行活体组织检查以诊断疾病。这一新系统破译了病变中生物标志物质的水平，为目前基于外观的评估补充了一致的、定量的信息。而且，真黑色素(一种赋予皮肤颜色的化学物质)和血红蛋白(红细胞中的蛋白质)的浓度和分布变化是黑色素瘤的强指标。

(编译自: https://www.sciencedaily.com/releases/2017/08/170823090930.htm)

(本刊讯)